White Paper Report

Report ID: 98487

Application Number: HD5050508

Project Director: Jennifer Golbeck (jgolbeck@umd.edu)

Institution: University of Maryland, College Park

Reporting Period: 8/1/2008-7/31/2009

Report Due: 10/31/2009

Date Submitted: 12/7/2010

ArchivesZ: Visualizing Archival Collections[*]

Jeanne Kramer-Smyth
College of Information Studies
University of Maryland
Hornbake Library, South Wing
College Park, MD 20742, USA
jkramers@mail.umd.edu

Morimichi Nishigaki
Computer Science Dept
University of Maryland
A.V. Williams Building
College Park, MD 20742, USA
michi@cs.umd.edu

Tim Anglade
Computer Science Dept
University of Maryland
A.V. Williams Building
College Park, MD 20742, USA
tag@umiacs.umd.edu

ABSTRACT

This paper proposes a method for the visualization and exploration of items associated with multi-value attributes for which there is overlap of attribute values across the data set. ArchivesZ is a prototype featuring this method and has been designed to support exploration of the metadata describing archival collections. Archival materials are unique and organized into groups, usually based on who created the materials. These groups may vary in size from a small number of folders to many hundreds of boxes. When describing archival materials, archivists note the range of years covered by the records, the size of the collection and the high level subjects of the records. One common metric for communicating the size of a collection is linear feet. A single linear foot is one foot of shelf space occupied by records. ArchivesZ leverages a unique dual sided histogram to support exploration of the multiple subjects assigned to each collection. It combines the dual sided histogram with a more traditional histogram displaying year data to permit tightly coupled, multi-dimensional browsing of subject and time period metadata. By representing the distribution of subjects and time periods using the metric of total aggregate linear feet of associated collections, ArchivesZ permits users to get a better sense of total available research materials than they would by viewing a standard search result list.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

Keywords

subject visualization, archives, search, EAD

1. INTRODUCTION

1.1 Subject Access and Archival Resources

The availablilty of centralized access to subject information about archival and manuscript collections is a fairly recent development. Archival practice, as developed in the United States in the middle of the twentieth century, revolves around the concepts of provenance and original order. The central concept is that archival records, because they are the product of the activities of an organization, should retain their original order when transfered into archival care. Background contextual information about records creators is crucial to understanding records. A side effect of grouping records by record creator and retaining the creator's original organization is that groups of records are described at the group level - not at the item level. Consider this in contrast with a library, where the central topic of a book usually dictates its location on the library shelf. The main access point to an archival record group or manuscript collection is the record creator. This basic difference between libraries and archives is key to understanding why subject access to archival resources is both challenging to achieve and very useful when available.

One of the major steps forward in establishing broad access to archives and manuscript collections was the National Union Catalog of Manuscript Collections (NUCMC). Published by the Library of Congress annually from 1962 through 1993, NUCMC ultimately provided indexing for the descriptions of over 70,000 collections - and included a subject index based on Library of Congress Subject Headings (LCSH).

The next major leap forward in centralized access to information about archival resources was triggered by the adoption of the MAchine-Readable Cataloguing (MARC) data standard by libraries. This is the standard which permitted the creation of computer based national level union catalogs of books held in libraries across the country. In 1983, the MARC Format for Archival and Manuscripts Control (MARC AMC) was approved by both the archival and library communities. MARC AMC established a standard by which an entire archival or manuscript collection could be described in a MARC record, and thereby included in the centralized union catalogs. While the MARC AMC record format captures a very small portion of the information archivists and manuscript curators provide to researchers via multi-page finding aids, they provided a way for researchers to discover that the resources in question existed. MARC AMC also encouraged the assignment of subjects to collec-

[*] ArchivesZ was created in partial fulfillment of the requirements of course CMSC734. A full version of this paper is available at `http://wiki.cs.umd.edu/cmsc734/`

tions - and encouraged that those subjects follow the LCSH standard for terminology. While this was a great challenge to archivists, accustomed up to this point to describing their archival collections in finding aid documents held only at the local repositories, it was a great boon to researchers. Up to the advents of NUCMC and MARC AMC, researchers usually depended on references to discover materials related to their topic of interest.

The standardization of collection description required by MARC AMC laid the foundation for the development of Encoded Archival Description (EAD), the current version of which was published in 2002. EAD is a Document Type Definition (DTD) defined in XML (Extensible Mark-up Language) that defines a standard for the encoding of finding aids for use online. This international de facto standard for encoding finding aids provides a springboard for the creation of software programs intended to extract, organize, facilitate discovery of and aggregate information about groups of archival resources at a much more granular level than was possible with MARC AMC. As Claire Gabriel concludes in "Subject Access to Archives and Manuscript Collections: An Historical Overview"[6]:

> Archivists have made vast progress toward providing improved subject access to archives and manuscript collections since the introduction of the MARC AMC format. This is all the more remarkable when one considers that throughout most of the twentieth century descriptive practices encouraged local creativity and few repositories had any means of unified access to collections. Both AMC and EAD have facilitated the goals of standardized description and improved searching capability that will promote knowledge and use of collections; the combination of these formats enables research across a broad spectrum of subject inquiry by a variety of users.

1.2 Opportunities for Visualization

Archives have lagged behind their library counterparts in their use of computers to support users" search activities. In order to support online access to metadata about their materials, libraries have the option to reuse existing catalog records. These catalog records include descriptive information including subject terms.

In contrast to the materials in libraries, archival materials are unique. Archival records and manuscripts are organized into groups, often called collections. Each collection requires the creation of a custom description - the finding aids mentioned in the prior section. Each finding aid presents background and contextual information to support understanding and use of the records. Finding aids include information about who created the records, when they were created, why they were created, what topics the records relate to and the size of the collection.

One of the greatest challenges to those who work with collections of archival records and manuscripts is understanding the quantity of available materials. The nature of archival materials is such that a single archival collection may consist of a single folder or hundreds of boxes. Researchers who usually use the standard library Online Public Access Catalog (OPAC) to find resources are accustomed to expecting one catalog listing to correspond with a single book or journal article. Viewing a standard search result list that just shows the title and short description of a collection makes it virtually impossible to get a handle on the quantity and subjects covered by each collection. While a list of ten library catalog records will usually correspond to ten books, a list of ten archival collections could represent anything from 10 small boxes to 10,000 boxes and anything in between. The size of the collection may be noted in many ways, but one common metric used is linear feet. The Society of American Archivists" glossary defines the linear foot as "A measure of shelf space necessary to store documents.'

There has been some research concerning how well internet search engines (such as Google and Excite) support search for archival finding aids[7]. It is appealing to imagine discovery of archival finding aids using Google. Unfortunately the lack of access to structured finding aid metadata can make the keyword style searching more frustrating than useful. Consider date ranges for example. If a researcher is trying to find records relating to 1954, but all the finding aids relating to relevant collections show date ranges that include 1954 but don't actually show the string 1954 in the text of the finding aid - none of these collections will be returned.

Chris Anderson of Wired described the power of the "long tail" in his Wired article of the same name. He discussed that the future belonged not to the bestsellers, but rather to "the millions of niche markets at the shallow end of the bit-stream."[1] There has been much discussion of the long tail with regard to library resources[3]. It is interesting to consider that when dealing with archival records, frequently everything is long tail. The nature of archival collections is such that many of those with the greatest desire to access the collections have very narrow and specific interests. It is quite rare that the documents in a single archival collection will be popular, in the sense of a bestselling book. Frequently it is a challenge for those wishing to use archival materials to figure out how to approach the search process. Use of a visualization tool, such as ArchivesZ, could support a more serendipitous process of exploration and discovery of relevant materials.

To support exploration of subject terms associated with collections, ArchivesZ leverages a unique dual sided histogram (see Figure 1). As subject terms are selected, the dual sided histogram chart is generated to display related subjects. One bar will be generated on the chart for each subject associated with collections returned based on the the selected subject term(s) and other search criteria. The size on the left side of the histogram will be determined by the total size of all collections returned based on the search criteria that share the selected attribute value and the charted attribute value. The size on the right side of the histogram will be determined by the size of all collections returned based on the search criteria that have the charted attribute value but do NOT have the selected attribute value. A secondary histogram bar is displayed behind the left half of the histogram to show the total size of all collections assigned only the selected subjects.

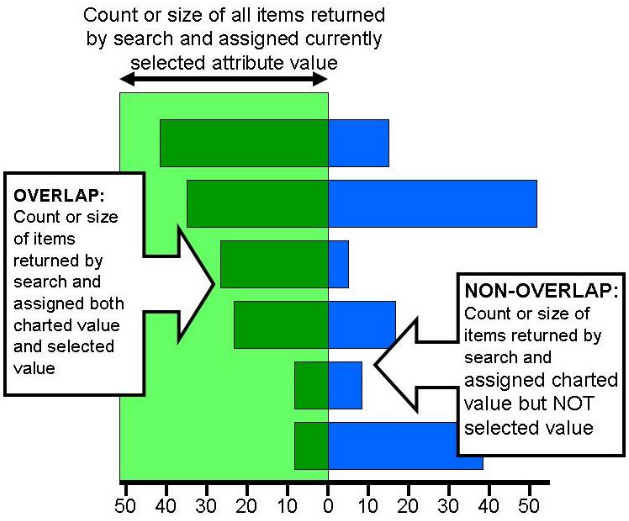

Count or size of all items returned by search and assigned currently selected attribute value
OVERLAP: Count or size of items returned by search and assigned both charted value and selected value
NON-OVERLAP: Count or size of items returned by search and assigned charted value but NOT selected value
50 40 30 20 10 0 10 20 30 40 50

Figure 1: Dual Sided Histogram

ArchivesZ combines the dual sided histogram with a more traditional histogram displaying year data to permit tightly coupled, multi-dimensional browsing of subject and time period metadata. By representing distribution of subjects and time periods using the metric of total aggregate linear feet of associated collections, ArchivesZ permits users to get a better sense of total available research materials than they would by viewing a standard search result list. The subject term visualization interface also supports a deeper understanding of related subject terms.

The interface presentation of year and subject term data is tightly coupled - as one dimension is manipulated, the other dimension is updated based on a refinement of collections returned.

1.3 Target Audience

The ArchivesZ prototype was designed to support the needs of three distinct user groups:

- Archivists
- Researchers
- Students

Archivists might use ArchivesZ to improve their understanding of the collections at various institutions including their own. They also may use the tool to ensure that the metadata currently associated with their collections matches the reality of what they know to be the case from hands on experience.

Researchers with very specific interests might use ArchivesZ to permit easy identification of institutions with archival collections fitting the criteria of their research. It is frequently the case that researchers must travel to archives in order to do their research[14], and a rapid grasp of the quantity of materials that cover the time period and subjects of interest may be an aid in planning.

ArchivesZ could enable exploration of locally held archival collections by students to promote use of primary materials. In contrast to researchers who frequently have very specific interests before they examine the collections held by an institution, students in the university setting likely are not aware of what primary sources are available. A tool like ArchivesZ might encourage the browsing and open ended exploration of the available collections.

1.4 Related Work

1.4.1 Existing Archival Collection Search Interfaces

There exist a number of portals for searching union catalogs of archival collection descriptions. Most of these are fee based services. ArchivesUSA provides a directory of 5,581 repositories and 160,792 collections of United States primary source material[1]. There are no visualization features available in the ArchivesUSA interface. ArchiveGrid.org is another fee based service that provides access to collection descriptions. While some collections are collected via a proprietary web crawler, most of the collection descriptions are based on catalog records in the RLG Union Catalog[2]. A great deal of effort is still being spent in figuring out the best methods of converting existing finding aids to EAD[11]. Often the most common method of archival collection discovery is via a library OPAC MARC record.

1.4.2 Multi-dimensional browsing interfaces

Traditional browse interfaces often only permit users to navigate a single hierarchical tree in order to discover items of interest. Multi-dimensional browsing shows more than one facet of the information simultaneously. These interfaces permit users to switch laterally between dimensions as they explore the data[13].

Perspectives Browser (PB) is designed to be a domain independent interface that uses parallel histograms to support exploration of multiple dimensions of attributes.

> Dual encoding of the histogram bars enables multivariate pattern discovery. Width shows the unconditional distribution over the whole collection, while height shows the conditional distribution given the current query.[4]

BungeeView is a prototype implementation of PB based on an image collection of historic Pittsburgh photographs[3].

Moritz Stefaner has developed a demonstration of the utility of “elastic lists” to support users exploration of multi-faceted data[4]. As a user changes the value of one attribute, the remaining attribute values displayed are adjusted accordingly.

Both the MetaCombine Project[5] and the Flamenco Search Project[6] has developed a method of displaying faceted search

[1] http://archives.chadwyck.com/

[2] http://www.archivegrid.org/web/jsp/h.jsp

[3] http://cityscape.inf.cs.cmu.edu/bungee/HistoricPittsburgh.html

[4] http://well-formed-data.net/archives/54/elastic-lists

[5] http://www.metacombine.org/

[6] http://flamenco.berkeley.edu/index.html

results. These tools do not present visualization of the records within each facet - but do an excellent job purely with text. The Relation Browsers developed at the Interaction Design Laboratory at UNC, Chapel Hill[10] and paperLens[9] provide tools to support user exploration of the relationships among various types of metadata. The recent work on generation of fast-list organization of search results provide rapid on the fly classification of web search results based on attributes determined from the URL itself[8].

One example of interesting work being done related to faceted search of digitized primary source materials is NINES. “NINES is a federation of peer-reviewed resources and innovative research tools, made freely available to students and scholars of 19th-century culture.”[7]. NINES has been built using Collex[8]. Collex is an open source tool developed by the Applied Research in Patacriticism at the University of Virginia. The interface provides users with realtime feedback across multiple facets as search criteria is selected and refined.

1.4.3 Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)

The Open Archives Initiative was developed with the intent of supporting increased scholarly access to e-print archives. While not aimed at harvesting metadata about archival collections or their items, there has been research into what it would take to generate OAI records for EAD encoded finding aids.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has great potential to enable the harvesting of archival collection metadata. The importance of collection level metadata has also been examined. Some work is being done to examine the efficacy of including collection level metadata as a way of preserving the context of items[12] being harvested based on the OAI-PMH.

1.4.4 Archon: Leveraging Subject Type

The recently launched website of the University of West Florida Libraries Holdings Database is based on an archives software project called Archon. Available free of charge for use by non-profit institutions, this tool is a ”web-based tool for archivists and manuscript curators. It automatically publishes archival descriptive information and digital archival objects to a user-friendly website.”[9]. Archon leverages the EAD support for subject types in their “Browse Subjects” view[10]. This gives a powerful method for users to explore subjects terms of various types (such as Corporate Name, Genre/Form of Material, Occupation or Topical Term). Drilling down into a specific subject term displays all collections assigned that subject term.

This does not take advantage of this structured data to display any visualizations of distribution of materials. It also does not support browsing by subject within a set of search results previously narrowed by a keyword search.

1.4.5 Collection Understanding

In their paper Collection Understanding, the authors document their development of an application that provides users with innovative approaches to exploring online collections of images. One example is a “streaming collage” that gives the users a big picture sense of the images available in a specific collection. They contrast their approaches, those permitting the users to gain a gestalt understanding of the types of images available, with the traditional approach of requiring explicit searches.

> Information retrieval (IR) is traditionally used as a tool for finding specific artifacts. Users must be able to define queries by specifying values for metadata fields. IR interfaces facilitate the “find the needle(s) in the haystack”' approach. Collection understanding is, in some sense, directly opposite to the IR approach. The users may have no prior knowledge of the metadata fields or values.[2]

ArchivesZ takes a similar approach to supporting users understanding not of a single collection, but rather permits the users to gain an impression of the types of collections and time periods available from one or more repositories.

2. ARCHIVESZ INTERFACE

2.1 Design Considerations

Driven by the desire to improve users” grasp of related subjects, many ideas for representing the relationship among collection subjects were considered. Node-network diagrams with an implementation of clustering was considered. This approach was passed over due to the difficulties related to ensuring visibility of all nodes. An “elastic lists” style approach was considered - but passed over due to the very large number of subjects (over 11,000 tags generated for the 802 collections included in our sample data-set). Hierarchical displays of subject terms were considered - but Library of Congress Subject Headings (LCSH) are not hierarchical. While there has been some work to attempt to map LCSH subjects to the Dewey Classification System (DCS)[11], and interesting visualizations built to explore the DCS hierarchy[12], much of this work stems from analysis of existing MARC records in large union catalogs of books and is very much a work in progress.

2.2 Interface Description

The final design of ArchivesZ leverages a dual sided histogram to communicate the intersection of subject terms assigned to collections.

The ArchivesZ interface supports visualization of aggregate information about groups of archival collections. The initial search permits the user to select a Repository, year range and keyword. The search results appear as shown in Figure 2. The bar chart on the left displays the range of decades covered by all collections returned by the search. The width of the bar corresponds with the total linear feet of all collections that contain at least one year within the decade, with

[7] http://nines.org/collex

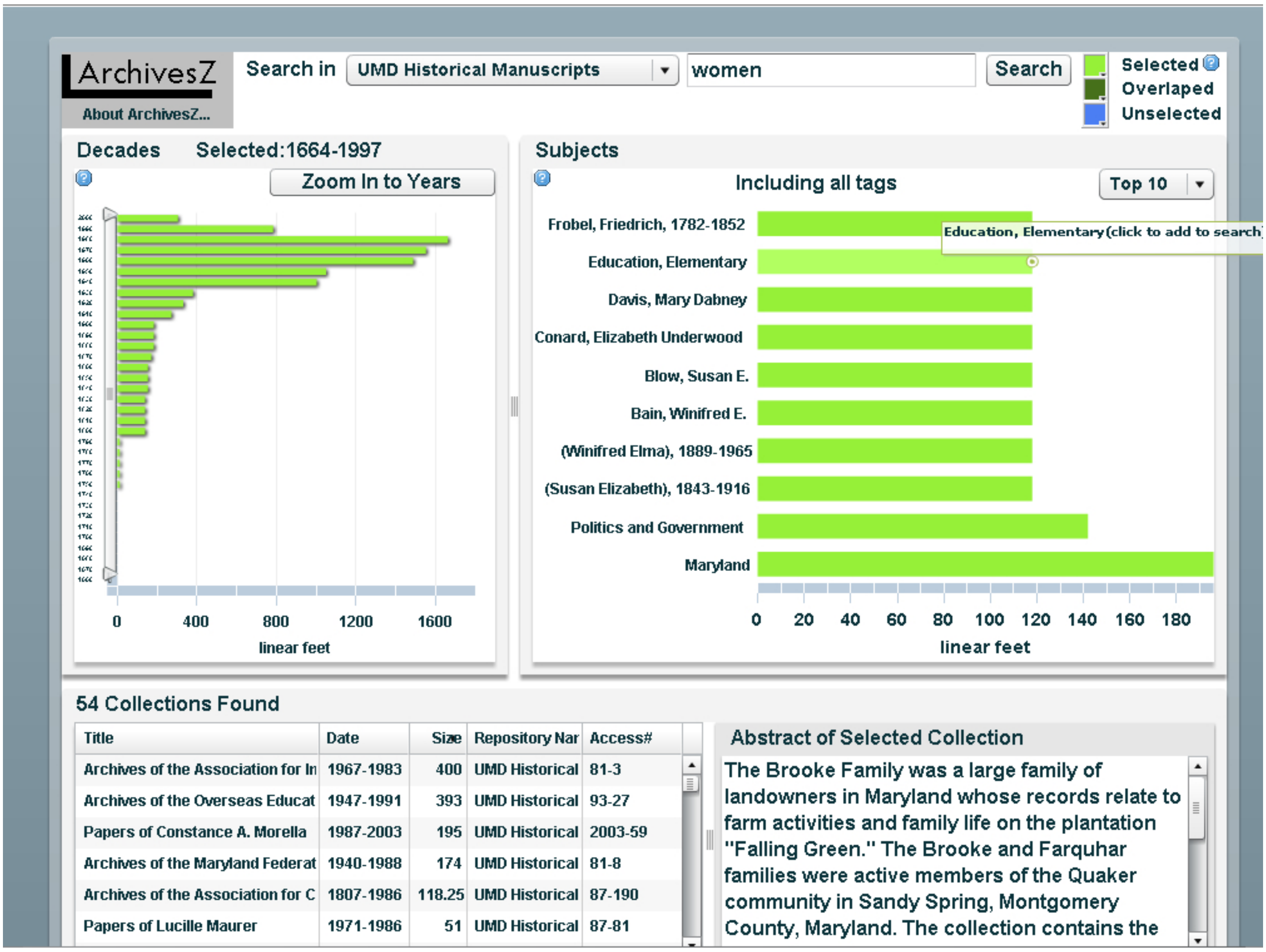
[8] http://www.nines.org/tools/collex.html

[9] http://www.archon.org

[10] http://fusionmx.lib.uwf.edu/archon/subjects.php

[11] http://www.oclc.org/dewey/updates/numbers/

[12] http://deweyresearch.oclc.org/ddcbrowser/wcat



Title	Date	Size	Repository Nar	Access#
Archives of the Association for In	1967-1983	400	UMD Historical	81-3
Archives of the Overseas Educat	1947-1991	393	UMD Historical	93-27
Papers of Constance A. Morella	1987-2003	195	UMD Historical	2003-59
Archives of the Maryland Federat	1940-1988	174	UMD Historical	81-8
Archives of the Association for C	1807-1986	118.25	UMD Historical	87-190
Papers of Lucille Maurer	1971-1986	51	UMD Historical	87-81

The Brooke Family was a large family of landowners in Maryland whose records relate to farm activities and family life on the plantation "Falling Green." The Brooke and Farquhar families were active members of the Quaker community in Sandy Spring, Montgomery County, Maryland. The collection contains the

Figure 2: ArchivesZ: First Level Search Results. Gives full overview of years and top 10 subjects by total linear feet.

each collection only adding their size to the total one time per decade. The bar chart on the right displays the top five subject terms based on the total linear feet worth of collections associated with that subject term. At the bottom of the screen can be found a list of the collections returned by the search.

At this point the user has several options for further exploration or refinement of the results. The left edge of the Decades histogram offers a double edged slider to permit refining of the decades included in the search. The [Zoom In to Years] button will switch the display to the years level rather than the decades level. Figure 3 shows what the years level display looks like after a user has refined their selection of years by using the sliders. The range of years may be limited or moved to change the collections returned as well as the subject terms displayed.

On the subject histogram, users may increase the number of displayed subjects. When the subject labels become too small to read, the user may view the name of the subject by placing their mouse over any bar. Tooltips were used throughout the application to provide help and permit browsing of attribute values when labels become too small.

If the user clicks on a subject bar - the application will add the subject term to the query criteria and resubmit the query. After the screen completes its refresh, the subjects panel will include a new set of information displayed using the dual sided histogram structure. The large green rectangle represents the selected subject term. In Figure 4 the selected subject term is Maryland. The dark green portions of the subject bars that are displayed within the Maryland selected term green box are referred to as "Overlap" subjects. The width of this bar represents how many linear feet of collections share both the selected term and the subject represented by the bar. The blue portions of the subject bars communicate the total size of all collections returned by the query that have the charted subject term but do NOT share the currently selected subject.

2.3 Universal Usability

ArchivesZ provides an easy method for users to select any three colors they prefer to correspond to the selected criteria, overlapping collections and non-overlapping collections. This ensures that users can choose optimum colors based on their own preferences and eyesight.

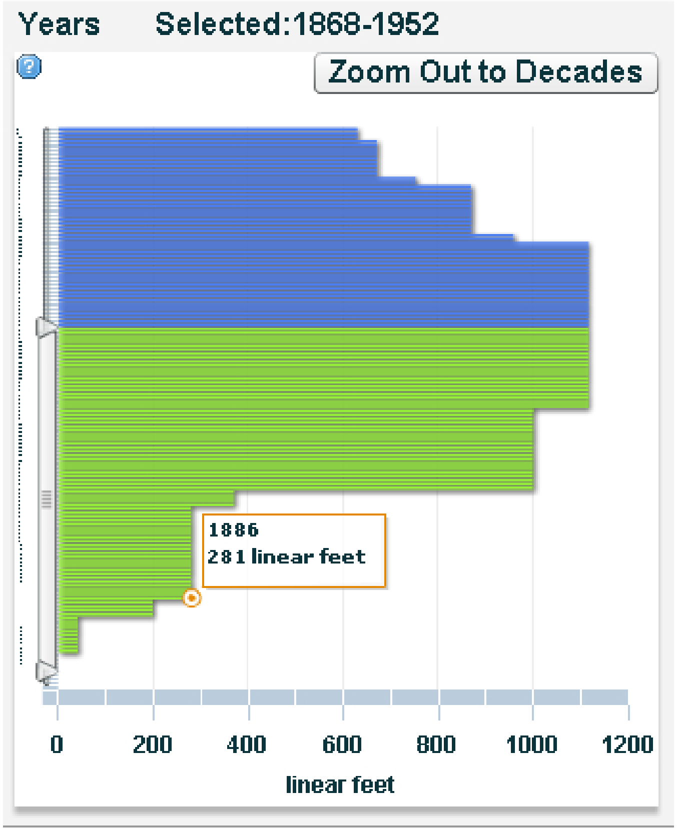


Figure 3: ArchivesZ Years Bar Chart: view after user has moved vertical sliders

3. DATA SOURCES AND MANIPULATION

3.1 Encoded Archival Description

Users searching for records in an archives typically have used the archival finding aid to assist in understanding the contents of individual collections. The adoption of the Encoded Archival Description (EAD) Document Type Description (DTD) as a standard by the archival community[13] has provided a source for standardized structured data about archival collections. ArchivesZ leverages the de facto standard of Encoded Archival Description (EAD) for the encoding archival finding aids in XML format. The application sought to leverage the aggregation of subject, year and collection size metadata to provide interactive visualizations of archival collections.

3.2 Data Sources

The data set used when building ArchivesZ included XML format EAD encoded finding aids from the University of Maryland Archives[14] as well as those publicly available from the Library of Congress[15].

In preparation for building ArchivesZ the Encoded Archival Description (EAD) Document Type Definition (DTD) was analyzed to isolate the structured data that would be useful

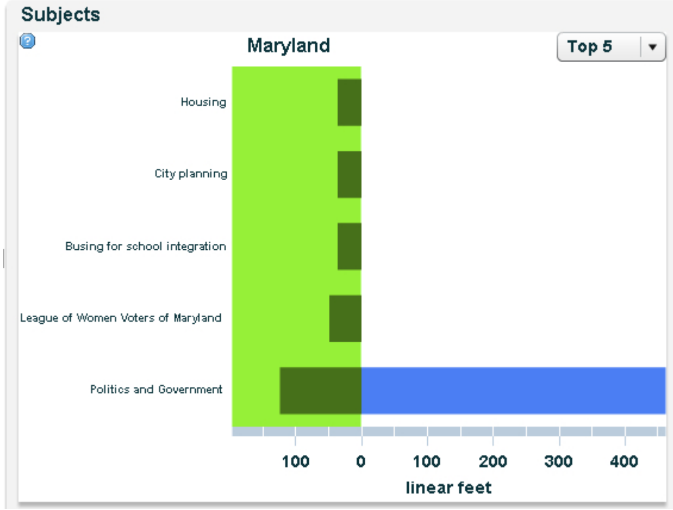


Figure 4: ArchivesZ Subject Bar Chart: view after user has selected the Maryland subject showing dual sided histogram display

for our visualization goals. A data model was created for use in building our database repository (Figure 5). A parser was created to permit extraction of the elements of interest from the XML files in order to populate the tables in the MySQL database. The parser program written in Ruby with REXML[16], which is an XML 1.0 conforming toolkit with intuitive API. While the data used in the ArchivesZ prototype depended upon data extracted from EAD finding aids, it would be possible to support the same visualizations of collections no matter what the data sources - as long as all the necessary elements were found to populate the required tables and columns.

3.2.1 Collection Size

While many collections do specify their size using linear feet, there are others that use other metrics. For example, a collection may only say that it has 300 microfilm reels or 135 photographs. Based on feedback from our adviser Archivist Jennie Levine, we used the following size conversion rules to convert all sizes into linear feet:

- 1 microfilm reel = 1 linear foot
- Collections represented only by a number of items will be represented as .25 linear feet
- If size only specified in number of boxes 1 box = .5 linear feet
- When the size of a single collection is given in multiple types of units, they are prioritized in the following order:
  1. linear feet
  2. boxes
  3. microfilm reels
  4. items

[13] http://www.archivists.org/saagroups/ead/aboutEAD.html
[14] http://www.lib.umd.edu/archivesum/index.jsp
[15] http://lcweb2.loc.gov/faid/source.html
[16] http://www.germane-software.com/software/rexml/

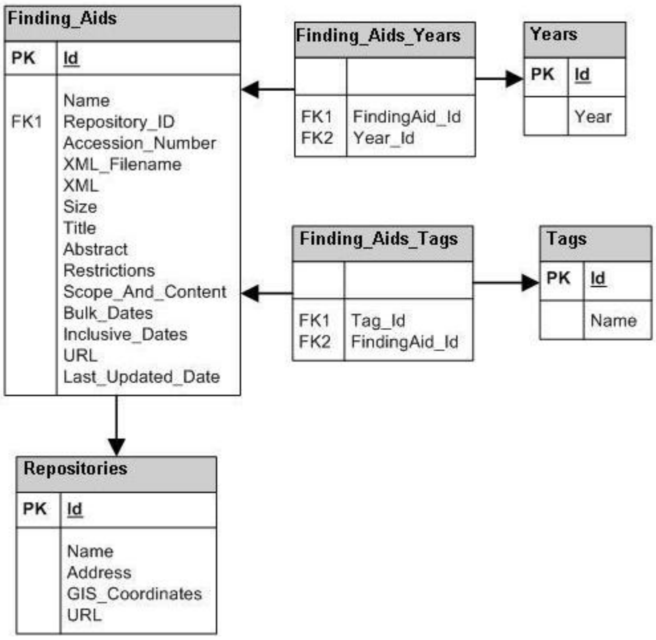


Figure 5: ArchivesZ Data Model

While these choices are not absolutely precise, they were adequate for the purposes of supporting our visualization of total linear feet of materials.

3.2.2 *Subjects and Tags*

In the EAD finding aids used for the ArchivesZ prototype, most subjects were found to be based on Library of Congress Subject Headings (LCSH). Due to the unique nature of each archival collection and the selection of very granular subjects, it would be unlikely to find a great deal of overlap among given LCSH subjects. In addition, we realized that the component parts of a standard LCSH subject would likely provide useful insight into the true aggregate nature of the collections.

For example, it should come as no surprise that many of the collections held in University of Maryland archives are topically related to the state of Maryland. Very specific LCSH subjects embed the term "Maryland" within them. Take the following two subjects:

- Agricultural colleges – Maryland – History – Sources
- Tobacco – Maryland – History – Sources

Using the full subjects as shown above would prevent the higher level understanding that collections with these two subjects are both concerning Maryland. To address this challenge we chose to break down the LCSH subjects into a list of what we termed "Tags". The list shown above would have resulted in the following tags:

- Agricultural colleges
- History
- Maryland
- Sources
- Tobacco

This choice resulted in the ability of ArchivesZ to associate aggregate collection size with the tag terms to present useful overviews of the archival collections.

Late in the development of the prototype, based on discussions with our adviser, Archivist Jennie Levine, we chose to remove those tags with which a very high percentage of collections were associated. We removed all associations with tags such as:

- Archives
- Correspondence
- History
- Sources

This cleanup of data after import into our database is consistent with what major projects utilizing the OAI-PMH have needed to do. There is an increasing understanding of the impact that metadata cleanup can have on the final usability of the data gathered. It remains to be seen if the responsibility for such cleanup lays with those harvesting data or those making records available for harvesting. As standards are established, being a good citizen of the shared data community will dictate that one can only benefit by following the standards[5].

It would be a very nice feature for a future version of ArchivesZ to permit local setting of "stop tags". This would support removal of terms that are not useful for local searches. If ArchivesZ were only used locally at the University of Maryland, it might make sense to remove the term "Maryland" from the tags used. That same tag might still be very useful if ArchivesZ were supporting the exploration of many archival collections at a national level.

For our data-set of 802 archival collection finding aids and their 12,604 associated subjects, we generated and retained 10,934 tags. 15 percent of the tags are associated with more than one collection. 102 of the tags (0.93 percent) are associated with more than 10 collections. This is in contrast with 10 percent of the subjects associated with more than one collection and 0.17 percent associated with more than 10 collections. A final determination of the best approach to handling subject terms cannot be made based on our small data sample, but these numbers show that our choice to simplify LCSH subjects into more atomic elements was a good choice for our data-set and the purposes of the ArchivesZ prototype.

3.3 Architecture

ArchivesZ is built on a software stack of Adobe Flex generated Flash, Ruby on Rails and a MySQL database (Figure 6). At runtime, the Flash application passes a set of search criteria to the Rails Web Server using an XML format. Subsequent requests to the Rails Web Server retrieve XML formated search results for each section of the Flex application. Separate requests retrieve a list of Finding Aids, Years and Tags data. This separation of data into distinct sets permits portions of the application to refresh on an as needed basis.

4. ARCHIVES EXPERTS AND STUDENTS

Jennie Levine, Curator for Historical Manuscripts University of Maryland Libraries, was our partner, domain expert and advisor for the ArchivesZ project. She supplied our team with over 500 EAD encoded finding aids in XML format. This served as the core of the data used when developing our application. She was in frequent contact with our team as we worked through resolving issues such as how to convert all collection sizes into linear feet.

We held an extensive feedback session with Ms. Levine near the conclusion of our project. As the Chair Elect of the Society of American Archivists EAD Roundtable, Ms. Levine has unique insight into ideas related to EAD. Ms. Levine spent an hour and a half session reviewing the ArchivesZ interface. Her feedback was very positive. She could envision multiple uses for the tool, including those outlined in our target audience section above.

The elements of the interface that Ms. Levine specifically identified as most useful were:

- the subject visualization
- the data slider
- the tooltips when used for exploring long lists of subjects (such as when the user has selected to view 100 subjects)

Elements that she identified as needing clarification or enhancement were:

- "stop" words for use with subjects (or some other way of eliminating common subject terms)
- ensuring that the dual sided histogram was not overwhelmed and made unreadable by a single large subject that reduces the scale of the "overlap area" to the point of it basically disappearing
- she expected that the keyword search included searching the subject terms, which it currently does not do

We discussed a number of the ways ArchivesZ could be used to support an archives. One that she seemed interested in was our notion of appealing to undergraduate university students on the local campus. She agreed that ArchivesZ could provide a "fun" way for students to get a quick, high level view of the quantity, time spans and subjects of the primary source materials available within the archives on campus.

Archivists rarely have time to re-examine finding aids after they have been written. She saw ArchivesZ as a tool to support the examination of the finding aids of her repository in an aggregate manner. We took it to be a very positive sign that we ran out of time in our demonstration before we ran out of questions that Ms. Levine wanted to answer using the interface.

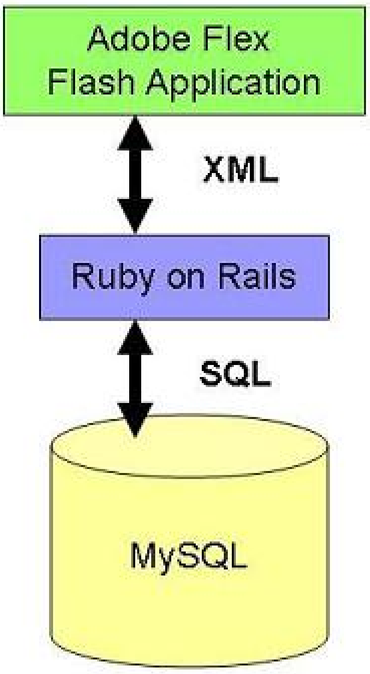


Figure 6: ArchivesZ Architecture

In addition to our session with Ms. Levine, ArchivesZ was shown as a demo in a graduate level course on Archival Access. The response from this audience was also enthusiastic. Students wanted to know when they might be able to try using the tool themselves. The instructor, Dr. Susan Davis, indicated a desire to ensure that the interface was easy enough for quick understanding.

5. CHALLENGES AND FUTURE WORK

5.1 Quality of Metadata

As has been mentioned above, libraries have a distinct advantage over archives with regards to metadata standardization. The unique nature of archives and the degree to which standardization of metadata values and use of subject authorities has evolved differently in many local archives around the world makes aggregating data challenging.

One of the interesting aspects of the feedback provided by our adviser was her desire to use ArchivesZ to examine the metadata applied to the collections in her local repository with an eye toward improving standardization.

5.2 Diversity of Subject Terms

Due to the unique nature of archival collections, the subject terms associated with the collections is very diverse. While it is certainly useful to view the top 5, 10, 20 or even 100 subjects by total collection size, as ArchivesZ now permits users to do - we believe that ultimately additional methods of exploring the "tail" subjects would be important in a future version of ArchivesZ.

5.3 Scaling to Support Large Data-sets

There are many minor interface issues that would need to be addressed in a final product intended to handle large data-sets. For example, currently selected subjects can be removed from the search criteria by unchecking a check box next to the listed subject term. When the number of selectable tags increase, a more scalable solution would need to be implemented. Due to chart component issue, the year labels are not readable when there are many years. While we currently work around this issue by showing years in data tip, we could reclaim some UI space by removing the illegible year labels altogether.

5.4 Query Performance

One of our ongoing challenges when building ArchivesZ was the optimization of the SQL queries used to return the search results for use by the Flex application to generate the visualizations. The most challenging of these was (and still is) the query that returns the details of the related tags. As of the writing of this paper, it often takes an unacceptable amount of time for the subject and related dual sided histogram to refresh.

One option for improving query performance is to use temporary tables to intermediately stage the results of the main query. This table of data could then be used as the user refines their selections of years and subjects.

Another option we consider very promising, but did not have time to implement involved returning each finding aid with their associated year and subject data to the Flex application. This would enable the logic for redrawing the subject tags to use local data rather than requiring a new request to the server. The only time a new set of data would need to be retrieved would be when the repository or keyword values were modified - likely to be perceived by users as beginning a new search.

5.5 Enhancement Ideas

5.5.1 Sparklines

While we are pleased with the ability of users to change the range of years selected on the Years barchart and see the corresponding change to subjects displayed, there are opportunities to improve visualization of this temporal relationship. Sparklines could be added to the subject bars to support greater understanding of the date ranges encompassed by collections assigned specific subject terms, as demonstrated in the Moritz Stefaner's Elastic Lists[17]. The term "Sparklines" was coined by Edward Tufte to describe "data-intense, design-simple, word-sized graphics". In our case each subject bar could display a small graphic that communicated the temporal distribution of records assigned the given subject term. If this visualization provided sufficient time period information, the Years/Decades panel could be hidden in favor of showing a full screen width or multipanel Subject panel. This additional screen space could be used for providing more subject filtering controls, which leads us to our next enhancement idea.

5.5.2 Subject Filtering

To support users exploration of all available subject terms, a few different methods could be employed. A horizontal slider could be added to the X axis of the Subject bar chart. This would permit users to remove the subject terms with the largest associated sizes in favor of those associated with fewer or smaller collections.

Another option would be to enable the user to create a specific list of subject terms to for display on the bar chart. The full list of subject terms could be made available for search to permit discovery and selection of terms to be displayed if the collections returned have been assigned these terms.

5.5.3 Search

Currently, keyword searching in ArchivesZ only string matches text in the Title, Scope and Content and Abstract of the finding aids. This was a conscious choice made by the ArchivesZ team in order to focus our attention on developing methods for visualizing the relationships among subject terms. The keyword search functionality should be extended to do a full text search of the entire finding aid. Due to the structured nature of much of the data in the finding aids - it would make good sense to provide an advanced search option to permit limitation of keyword search to specific portions of the finding aid such as biographical or subject data. The interface should also be modified to permit cross-repository and multi-repository search.

5.5.4 Geographical Visualization

Due to the association of each archival collection with a physical repository, it would be possible to generate geographical heat maps of the distribution of collections matching search criteria. This type of visualization could help researchers identify geographic areas and sets of archives that may be most productive to target.

6. CONCLUSIONS

We have shown our method for the visualization and exploration of items having multi-value attributes for which there is overlap of the attribute values assigned across the item data set. Through the creation of the ArchivesZ prototype, we have demonstrated the usefulness of providing this type of visualization.

ArchiveZ visualizes the overlapping assignment of subjects terms to archival collections. By leveraging the combination of key structured data elements of metadata about archival collections, the ArchivesZ provides end users with a way to correlate the size of collections with both the time and subjects covered. From an archives perspective, ArchivesZ adds in the missing metric of collection size when doing archival research. Nothing will replace the need to eventually read the finding aid of a collection of interest - but putting visualization tools in the hands of archives users will facilitate the understanding of the big picture of the materials available at a specific archives.

The dual sided histogram approach could be re-implemented to support exploration of any multi-variable data-set for which there is reuse of values across multiple items. The most classic example exposed on the internet today is the

[17] http://well-formed-data.net/archives/54/elastic-lists

assignment of unrestricted "tags" in applications such as Flickr.com and del.icio.us. Most approaches to examining the overlap in usage of tags has been using node-network diagrams or simple scoped lists. Our visualization could be re-implemented to support exploration and understanding of the relationships among tags.

7. ACKNOWLEDGMENTS

Thank you to our adviser Jennie Levine, Curator for Historical Manuscripts University of Maryland Libraries, for her support, insights and rapid e-mail answers to our many questions. Thank you to Dr. Susan Davis and the class of Spring 2007 LBSC 781 for taking the time to view a demonstration of ArchivesZ. This project and paper were created in partial fulfillment of the requirements of University of Maryland, College Park course CMSC734.

8. REFERENCES

[1] C. Anderson. The long tail. *Wired*, 12(10), 2004.

[2] M. Chang, J. J. Leggett, R. Furuta, A. Kerne, J. P. Williams, S. A. Burns, and R. G. Bias. Collection understanding. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 334–342, New York, NY, USA, 2004. ACM Press.

[3] L. Dempsey. Libraries and the long tail: Some thoughts about libraries in a network age. *D-Lib Magazine*, 12(4), April 2006.

[4] M. Derthick and J. Zimmerman. The perspectives browser: Exploratory data analysis for everyone. Submitted to the 2005 IEEE Symposium on Information Visualization, Carnegie-Mellon University and Human-Computer Interaction Institute, http://www.cs.cmu.edu/ sage/PDF/DerthickZimmerman.pdf, 2005.

[5] M. Foulonneau and T. W. Cole. Strategies for reprocessing aggregated metadata. In *Research and Advanced Technology for Digital Libraries: 9th European Conference, ECDL 2005 Proceedings.* Springer Berlin / Heidelberg, 2005.

[6] C. Gabriel. Subject access to archives and manuscript collections: An historical overview. *Journal of Archival Organization*, 1(4), 2002.

[7] P. E. Hymas. Can you find me now?: Re-examining search engines capability to retrieve finding aids on the world wide web. Master's thesis, University of North Carolina at Chapel Hill, School of Information and Library Science, 2005.

[8] B. Kules, J. Kustanowitz, and B. Shneiderman. Categorizing web search results into meaningful and stable categories using fast-feature techniques. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 210–219, New York, NY, USA, 2006. ACM Press.

[9] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using paperlens. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1969–1972, New York, NY, USA, 2005. ACM Press.

[10] G. Marchionini and B. Brunk. Towards a general relation browser: A gui for information architects. *Journal of Digital Information*, 4(1), 2003.

[11] D. Meissner. First things first: Reengineering finding aids for implementation of ead. *The American Archivist*, 60(4), Fall 1997.

[12] C. J. Prom and T. G. Habing. Using the open archives initiative protocols with ead. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 171–180, New York, NY, USA, 2002. ACM Press.

[13] R. Shen, N. S. Vemuri, W. Fan, R. da S. Torres, and E. A. Fox. Exploring digital libraries: integrating browsing, searching, and visualization. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10, New York, NY, USA, 2006. ACM Press.

[14] C. L. P. William S. Brockman, Laura Neumann and T. J. Tidline. *Scholarly Work in the Humanities and the Evolving Information Environment.* Council on Library and Information Resources, 2001.